CS 189/289A Introduction to Machine Learning
Spring 2024 Jonathan Shewchuk

HW2: I ♥ Math

Due Wednesday, February 7 at 11:59 pm

- Homework 2 is an entirely written assignment; no coding involved.
- We prefer that you typeset your answers using LaTeX or other word processing software. If you haven't yet learned LaTeX, one of the crown jewels of computer science, now is a good time! Neatly handwritten and scanned solutions will also be accepted.
- In all of the questions, **show your work**, not just the final answer.
- **Start early. This is a long assignment. Most of the material is prerequisite material not covered in lecture; you are responsible for finding resources to understand it.**

Deliverables:

1. Submit a PDF of your homework to the Gradescope assignment entitled "HW2 Write-Up". You may typeset your homework in LaTeX or Word (submit PDF format, **not** .doc/.docx format) or submit neatly handwritten and scanned solutions. **Please start each question on a new page.** If there are graphs, include those graphs in the correct sections. **Do not** put them in an appendix. We need each solution to be self-contained on pages of its own.
   - In your write-up, please state whom you had discussions with (not counting course staff) about the homework contents.
   - In your write-up, please copy the following statement and sign your signature next to it. (Mac Preview, PDF Expert, and FoxIt PDF Reader, among others, have tools to let you sign a PDF file.) We want to make it *extra* clear so that no one inadvertently cheats.

     "I certify that all solutions are entirely in my own words and that I have not looked at another student's solutions. I have given credit to all external sources I consulted."

1 Identities and Inequalities with Expectation

For this exercise, the following identity might be useful: for a probability event A, $\mathbb{P}(A) = \mathbb{E}[\mathbf{1}\{A\}]$, where $\mathbf{1}\{\cdot\}$ is the indicator function.

1. For any non-negative real-valued random variable X and constant $t > 0$, show that

$$\mathbb{P}(X \geq t) \leq \mathbb{E}[X]/t.$$

This is known as Markov's inequality.
Hint: show that for $a, b > 0$, $\mathbf{1}\{a \geq b\} \leq a/b$.

Solution: When $X \geq t$, $\mathbf{1}\{X \geq t\} = 1 \leq X/t$. On the other hand, when $X < t$, $\mathbf{1}\{X \geq t\} = 0 \leq X/t$ since X is non-negative. Thus $\mathbb{P}(X \geq t) = \mathbb{E}[\mathbf{1}\{X \geq t\}] \leq \mathbb{E}[X/t] = \mathbb{E}[X]/t$.

[RUBRIC: A completely correct solution gets **(+1 point)**.]

2. A common use for concentration inequalities (like Markov's inequality above) is to study the performance of a statistical estimator. For example, given a random real-valued variable Z with true mean μ and variance 1, we hope to estimate its mean using the regular estimator $\hat{\mu} = \sum_{i=1}^{n} Z_i/n$, where $\{Z_i\}_{1,\ldots,n}$ are n Independent and identically distributed (i.i.d.) real values sampled from the distribution. Using Markov's inequality above, please show that

$$\mathbb{P}(|\hat{\mu} - \mu| \geq t) \leq \frac{1}{\sqrt{n}t}.$$

This tells us that the more samples we have, the better our estimator will be.
Hint: Use reverse Jensen's inequality to convert the mean of an absolute value into a mean of squares.

Solution: Note that

$$\mathbb{P}(|\hat{\mu} - \mu| \geq t) \leq \frac{\mathbb{E}[|\hat{\mu} - \mu|]}{t} \leq \frac{\mathbb{E}[(\hat{\mu} - \mu)^2]^{1/2}}{t}$$

where the first inequality follows from Markov's inequality and the second follows from Jensen's. Then, making $\tilde{Z}_i = Z_i/n$, we have

$$\mathbb{E}[(\hat{\mu} - \mu)^2] = \mathrm{Var}\left[\sum_{i=1}^{n} \tilde{Z}_i\right] = \sum_{i=1}^{n} \mathrm{Var}[\tilde{Z}_i] = n\frac{\mathrm{Var}[Z_i]}{n^2} = 1/n$$

where the second identity follows from the additivity of variance of independent random variables. Combining this with the first equation gives

$$\mathbb{P}(|\hat{\mu} - \mu| \geq t) \leq \frac{\mathbb{E}[(\hat{\mu} - \mu)^2]^{1/2}}{t} = \frac{1}{\sqrt{n}t}$$

[RUBRIC: There could be other ways to solve this. Any completely correct solution gets **(+2 points)**. Any partially correct or incomplete solution gets **(+1 point)**.]

2 Probability Potpourri

1. Recall that the covariance of two random variables X and Y is defined to be $\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$. For a multivariate random variable Z (i.e., each component of the vector Z is a random variable), we define the square covariance matrix Σ with entries $\Sigma_{ij} = \text{Cov}(Z_i, Z_j)$. Concisely, $\Sigma = \mathbb{E}[(Z - \mu)(Z - \mu)^\top]$, where μ is the mean value of the (column) vector Z. Show that the covariance matrix is always positive semidefinite (PSD). You can use the definition of PSD matrices in Q3.2.

   Solution: For $v \in \mathbb{R}^n$, $v^\top \mathbb{E}[(X - \mu)(X - \mu)^\top]v = \Sigma_{i=1}^n \Sigma_{j=1}^n \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)]v_i v_j = \mathbb{E}[v^\top (X - \mu)(X - \mu)^\top v] = \mathbb{E}[((X - \mu)^\top v)^2] \geq 0$. Note that the second identity comes from linearity of expectation.

   [RUBRIC: A completely correct solution gets **(+1 point)**.]

2. The probability that an archer hits her target when it is windy is 0.4; when it is not windy, her probability of hitting the target is 0.7. On any shot, the probability of a gust of wind is 0.3. Find the probability that

   (i) on a given shot there is a gust of wind and she hits her target.
   (ii) she hits the target with her first shot.
   (iii) she hits the target exactly once in two shots.
   (iv) on an occasion when she missed, there was no gust of wind.

   Solution: Denote with H the event that she hits her target, and with W the event that there is a gust of wind. Then we know that: $P(H \mid W) = 0.4$, $P(H \mid W^c) = 0.7$ and $P(W) = 0.3$.

   (i) $P(H \cap W) = P(H \mid W)P(W) = 0.12$
   (ii) $P(H) = P(H \mid W)P(W) + P(H \mid W^c)P(W^c) = 0.61$
   (iii) This probability is $\binom{2}{1}P(H)P(H^c) = 0.4758$
   (iv) $P(W^c \mid H^c) = \frac{P(H^c|W^c)P(W^c)}{P(H^c)} = 0.538$

   [RUBRIC: A correct derivation & answer to a sub-part gets **(+0.5 point)**. Total **(+2 points)**.]

3. An archery target is made of 3 concentric circles of radii $1/\sqrt{3}$, 1 and $\sqrt{3}$ feet. Arrows striking within the inner circle are awarded 4 points, arrows within the middle ring are awarded 3 points, and arrows within the outer ring are awarded 2 points. Shots outside the target are awarded 0 points.

   Consider a random variable X, the distance of the strike from the center in feet, and let the probability density function of X be

$$f(x) = \begin{cases} \frac{2}{\pi(1+x^2)} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

What is the expected value of the score of a single strike?

Solution: The expected value is

$$
\begin{aligned}
& \int_0^{1/\sqrt{3}} 4\frac{2}{\pi(1+x^2)}\,\mathrm{d}x + \int_{1/\sqrt{3}}^{1} 3\frac{2}{\pi(1+x^2)}\,\mathrm{d}x + \int_1^{\sqrt{3}} 2\frac{2}{\pi(1+x^2)}\,\mathrm{d}x \\
= \ & \frac{2}{\pi}\left[4\left(\arctan\frac{1}{\sqrt{3}} - \arctan 0\right) + 3\left(\arctan 1 - \arctan\frac{1}{\sqrt{3}}\right) + 2\left(\arctan\sqrt{3} - \arctan 1\right)\right] \\
= \ & \frac{13}{6}.
\end{aligned}
$$

[RUBRIC: A correct derivation and answer gets **(+1 point)**.]

3 Linear Algebra Review

1. First we review some basic concepts of rank. Recall that elementary matrix operations do not change a matrix's rank. Let $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times p}$. Let I_n denote the $n \times n$ identity matrix.

   (a) Perform elementary row and column operations[1] to transform $\begin{bmatrix} I_n & 0 \\ 0 & AB \end{bmatrix}$ to $\begin{bmatrix} B & I_n \\ 0 & A \end{bmatrix}$.

   (b) Let's find lower and upper bounds on $\operatorname{rank}(AB)$. Use part (a) to prove that $\operatorname{rank} A + \operatorname{rank} B - n \le \operatorname{rank}(AB)$. Then use what you know about the relationship between the column space (range) and/or row space of AB and the column/row spaces for A and B to argue that $\operatorname{rank}(AB) \le \min\{\operatorname{rank} A, \operatorname{rank} B\}$.

   (c) If a matrix A has rank r, then some $r \times r$ submatrix M of A has a nonzero determinant. Use this fact to show the standard facts that the dimension of A's column space is at least r, and the dimension of A's row space is at least r. (Note: You will not receive credit for other ways of showing those same things.)

   (d) It is a fact that $\operatorname{rank}(A^\top A) = \operatorname{rank} A$; here's one way to see that. We've already seen in part (b) that $\operatorname{rank}(A^\top A) \le \operatorname{rank} A$. Suppose that $\operatorname{rank}(A^\top A)$ were strictly less than $\operatorname{rank} A$. What would that tell us about the relationship between the column space of A and the null space of $A^\top$? What standard fact about the fundamental subspaces of A says that relationship is impossible?

   (e) Given a set of vectors $S \subseteq \mathbb{R}^n$, let $AS = \{Av : v \in s\}$ denote the subset of $\mathbb{R}^m$ found by applying A to every vector in S. In terms of the ideas of the column space (range) and row space of A: What is $A\mathbb{R}^n$, and why? (Hint: what are the definitions of column space and row space?) What is $A^\top A\mathbb{R}^n$, and why? (Your answer to the latter should be purely in terms of the fundamental subspaces of A itself, not in terms of the fundamental subspaces of $A^\top A$.)

Solution:

(a) The operations are as follows.

$$\begin{bmatrix} I_n & 0 \\ 0 & AB \end{bmatrix} \Rightarrow \begin{bmatrix} I_n & 0 \\ A & AB \end{bmatrix} \Rightarrow \begin{bmatrix} I_n & -B \\ A & 0 \end{bmatrix} \Rightarrow \begin{bmatrix} B & I_n \\ 0 & A \end{bmatrix}$$

In order,

- Left multiply the first row by A and add it to the second row
- Right multiply the first column by B and subtract it from the second column
- Negate the right column and exchange it with the left one

[RUBRIC: A complete and correct set of of operations gets **(+1 point)**. A partially correct or incomplete set of operations gets **(+0.5 point)**.]

[1] If you're not familiar with these, https://stattrek.com/matrix-algebra/elementary-operations is a decent introduction.

(b) Using (a) and rearranging gives us the lower bound,

$$n + \operatorname{rank}(AB) = \operatorname{rank}\left(\begin{bmatrix} I_n & 0 \\ 0 & AB \end{bmatrix}\right) = \operatorname{rank}\left(\begin{bmatrix} B & I_n \\ 0 & A \end{bmatrix}\right) \geq \operatorname{rank} A + \operatorname{rank} B$$

[RUBRIC: Correctly proving the lower bound gets **(+1 point)**.]

Let $\mathcal{R}(M)$ denote the range (column space) of a matrix M. Since $\mathcal{R}(AB) \subseteq \mathcal{R}(A)$, we have $\operatorname{rank}(AB) \leq \operatorname{rank} A$. Similarly, since $\mathcal{R}(B^\top A^\top) \subseteq \mathcal{R}(B^\top)$—that is, the row space of AB is a subset of the row space of B—we have $\operatorname{rank}(AB) \leq \operatorname{rank} B$. Thus $\operatorname{rank}(AB) \leq \min\{\operatorname{rank} A, \operatorname{rank} B\}$.

[RUBRIC: Correctly proving the upper bound gets **(+1 point)**.]

(c) As M has a nonzero determinant, its columns are linearly independent, and so are its rows; so its column space and row space are both $\mathbb{R}^r$.

Therefore, the r columns of A from which M is drawn are linearly independent—as we go from M to A, lengthening the columns can't break their linear independence. Likewise, the r rows of A from which M is drawn are linearly independent. The column space of A is the set of vectors in $\mathbb{R}^m$ spanned by the columns of A, so it must have dimension at least r. The row space of A is the set of vectors in $\mathbb{R}^n$ spanned by the rows of A, so it must have dimension at least r.

[RUBRIC: **(+1 point)**, or **(+0.5 point)** for a nice try.]

(d) If $\operatorname{rank}(A^\top A) < \operatorname{rank} A$, then the column space of A has dimension $\operatorname{rank} A$, but when we multiply $A^\top$ by that column space, we get a subspace of smaller dimension. Therefore, there is a nonzero vector v in the column space of A that is also in the null space of $A^\top$. (In other words, the intersection of the column space of A with the null space of $A^\top$ contains more than just the origin.) But the column space of A is the row space of $A^\top$, which is well known to be orthogonal to the null space of $A^\top$. This implies that v is orthogonal to itself but nonzero, a contradiction.

[RUBRIC: **(+1 point)**, or **(+0.5 point)** for a nice try.]

(e) $A\mathbb{R}^n$ is the column space of A, because if A_{*i} denotes the ith column of A, then $Av = \sum_{i=1}^n v_i A_{*i}$, and as v ranges over $\mathbb{R}^n$, the set $A\mathbb{R}^n$ contains all possible linear combinations of the columns of A.

[RUBRIC: **(+1 point)** for explaining why $A\mathbb{R}^n$ is the column space of A.]

$A^\top A\mathbb{R}^n$ is the row space of A, because $\mathcal{R}(A^\top A) \subseteq \mathcal{R}(A^\top)$, but it can't be a strict subset, because $\operatorname{rank}(A^\top A) = \operatorname{rank} A$. $\mathcal{R}(A^\top)$ is the row space of A.

[RUBRIC: **(+1 point)** for explaining why $A^\top A\mathbb{R}^n$ is the row space of A.]

[RUBRIC: **Total (+7 points)**.]

2. Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix. Prove equivalence between these three different definitions of positive semidefiniteness (PSD). Note that when we talk about PSD matrices, they are defined to be symmetric matrices. There are nonsymmetric matrices that exhibit PSD properties, like the first definition below, but not all three.

(a) For all $x \in \mathbb{R}^n$, $x^\top Ax \geq 0$.

(b) All the eigenvalues of A are nonnegative.

(c) There exists a matrix $U \in \mathbb{R}^{n\times n}$ such that $A = UU^\top$.

Positive semidefiniteness will be denoted as $A \succeq 0$.

Solution: (a) $\Rightarrow$ (b): Let λ be an eigenvalue of A with corresponding eigenvector v. Then

$$v^\top A v = \lambda v^\top v = \lambda \|v\|^2.$$

By part (a), we know that $\lambda\|v\|^2 \geq 0$, so $\lambda \geq 0$.
(b) $\Rightarrow$ (c): Consider the eigendecomposition of A, $A = V\Lambda V^\top$, where Λ is a diagonal matrix with entries equal to the eigenvalues of A, $\lambda_1, \ldots, \lambda_n$. Define $U := V\sqrt{\Lambda}$, where $\sqrt{\Lambda}$ is diagonal with entries equal to $\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}$; notice that this choice is justified because, by assumption, the eigenvalues are non-negative. Clearly, $A = UU^\top$.
(c) $\Rightarrow$ (a): Let $x \in \mathbb{R}^n$. Then

$$x^\top A x = x^\top UU^\top x = (U^\top x)^\top (U^\top x) = \|U^\top x\|^2 \geq 0.$$

[RUBRIC: Correctly proving any of the required three directions for equivalence gets **(+1 point)**.]

[RUBRIC: **Total (+3 points)**.]

3. The Frobenius inner product between two matrices of the same dimensions $A, B \in \mathbb{R}^{m\times n}$ is

$$\langle A, B\rangle = \operatorname{trace}(A^\top B) = \sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}B_{ij},$$

where trace M denotes the *trace* of M, which you should look up if you don't already know it. (The norm is sometimes written $\langle A, B\rangle_F$ to be clear.) The Frobenius norm of a matrix is

$$\|A\|_F = \sqrt{\langle A, A\rangle} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |A_{ij}|^2}.$$

Prove the following. The Cauchy–Schwarz inequality, the cyclic property of the trace, and the definitions in part 2 above may be helpful to you.

(a) $x^\top A y = \langle A, xy^\top\rangle$ for all $x \in \mathbb{R}^m, y \in \mathbb{R}^n, A \in \mathbb{R}^{m\times n}$.

(b) If A and B are symmetric PSD matrices, then $\operatorname{trace}(AB) \geq 0$.

(c) If $A, B \in \mathbb{R}^{n\times n}$ are real symmetric matrices with $\lambda_{\max}(A) \geq 0$ and B being PSD, then $\langle A, B\rangle \leq \sqrt{n}\lambda_{\max}(A)\|B\|_F$.
Hint: Construct a PSD matrix using $\lambda_{\max}(A)$

Solution:

(a) Realize that $x^\top A y = \langle A, xy^\top\rangle = \sum_{i=1,j=1}^{m,n} x_i * A_{ij} * y_j$. [RUBRIC: Stating the above equation or anything equivalent gets **(+1 point)**.]

(b) By the third definition of PSD, let $A = UU^\top$ and $B = VV^\top$. Then

$$\text{trace}\,(AB) = \text{trace}\,(UU^\top VV^\top) = \text{trace}\,(U^\top VV^\top U) = \text{trace}\,(U^\top V(U^\top V)^\top) \geq 0,$$

which follows because $M \stackrel{\text{def}}{=} U^\top V(U^\top V)^\top$ is PSD by the third definition, and trace $M \geq 0$, since trace is the sum of all eigenvalues. [RUBRIC: Writing trace (AB) as a trace of PSD matrix gets **(+0.5 point)**. Proving that trace $M \geq 0$ gets **(+0.5 point)**.]

(c) First realize that $\lambda_{\max}(A)I_n - A \geq 0$. This follows from the (a) definition of PSD matrices. Then by the identity proved above, trace $((\lambda_{\max}(A)I_n - A)B) \geq 0$, and rearranging gives trace $(AB) \leq \lambda_{\max}(A)$trace (I_n, B). By the second identity proved in this question we have trace $(I_n, B) \leq \sqrt{n}\|B\|_F$. [RUBRIC: Realizing trace $((\lambda_{\max}(A)I_n - A)B)$ gets **(+1 point)**. Completing other parts of the solution gets **(+1 point)**.]

[RUBRIC: **Total (+4 points)**.]

4. Let $A \in \mathbb{R}^{m\times n}$ be an arbitrary matrix. The maximum singular value of A is defined to be $\sigma_{\max}(A) = \sqrt{\lambda_{\max}(A^\top A)} = \sqrt{\lambda_{\max}(AA^\top)}$. Prove that

$$\sigma_{\max}(A) = \max_{\substack{u\in\mathbb{R}^m,\ v\in\mathbb{R}^n \\ \|u\|=1,\ \|v\|=1}} (u^\top Av).$$

Solution: In this solution, we abbreviate $\sigma_{\max}(A)$ as σ for notational convenience.

The proof will decompose into two steps. We will first show that σ^2 can be written as the supremum of a quadratic form over the unit sphere of $\mathbb{R}^n$. We will then reformulate this supremum to show the desired result.

Step 1: Observe that A^TA is a symmetric matrix, and that $\forall x \in \mathbb{R}^n$, we have $x^TA^TAx = (Ax)^T(Ax) = \|Ax\|^2 \geq 0$, that is, A^TA is positive semi-definite. We can, therefore, apply the spectral theorem to A^TA. There exists an orthogonal matrix $U \in \mathbb{R}^{n\times n}$, and a diagonal matrix $\Lambda \in \mathbb{R}^{n\times n}$ such that $A^TA = U\Lambda U^T$. Consequently,

$$\sup_{v,\|v\|=1} v^TA^TAv = \sup_{v,\|v\|=1} v^TU\Lambda U^Tv.$$

Consider the change of variable $w = U^Tv$. Since U^T is orthogonal, the image of the unit sphere $\{v \in \mathbb{R}^n, \|v\| = 1\}$ under the transformation $v \mapsto U^Tv$ is the unit sphere $\{w \in \mathbb{R}^n, \|w\| = 1\}$. Therefore,

$$\sup_{v,\|v\|=1} v^TA^TAv = \sup_{w,\|w\|=1} w^T\Lambda w.$$

Observe that $w^T\Lambda w = \sum_{i=1}^n \lambda_i w_i^2$, where λ_i is the i-th largest eigenvalue of A^TA. The supremum of this sum is achieved when w is the unit vector associated with the largest eigenvalue of A^TA, that is, σ^2. This implies that $\sigma^2 = \sup_{w,\|w\|=1} w^TA^TAw$.

Step 2: In parallel,

$$\sup_{v,\|v\|=1} v^T A^T A v = \sup_{v,\|v\|=1} \|Av\|^2$$

The application of the Cauchy-Schwarz theorem yields that for any unit vector $u \in \mathbb{R}^n$, $u \cdot Av \leq \|Av\|\|u\| = \|Av\|$, and that this inequality is tight when u is a scalar multiple of Av. This implies that $\|Av\| = \sup_{u,\|u\|=1} u^T Av$, and therefore,

$$\sigma^2 = \sup_{v,\|v\|=1} \sup_{u,\|u\|=1} \left[u^T Av\right]^2,$$

which concludes the proof.

[RUBRIC: **Total (+2 points)**.]

4 Matrix/Vector Calculus and Norms

1. Consider a 2×2 matrix A, written in full as $\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$, and two arbitrary 2-dimensional vectors x, y. Calculate the gradient of

$$\sin(A_{11}^2 + e^{A_{11}+A_{22}}) + x^\top Ay$$

with respect to the matrix A.
Hint: The gradient has the same dimensions as A. Use the chain rule.

Solution: We will take derivatives of $\sin(A_{11}^2 + e^{A_{11}+A_{22}})$ and $x^\top Ay$ separately.

First consider the former term, and take the derivative with respect to A_{11}.

$$\frac{\partial}{\partial A_{11}} \sin(A_{11}^2 + e^{A_{11}+A_{22}}) = \cos(A_{11}^2 + e^{A_{11}+A_{22}})[(2A_{11}) + e^{A_{11}+A_{22}}).$$

A similar computation can be done for A_{22}. Therefore, the derivative of $\sin(A_{11}^2 + e^{A_{11}+A_{22}})$ with respect to A is

$$\begin{bmatrix} \cos(A_{11}^2 + e^{A_{11}+A_{22}})[(2A_{11}) + e^{A_{11}+A_{22}})] & 0 \\ 0 & \cos(A_{11}^2 + e^{A_{11}+A_{22}}) \cdot e^{A_{11}+A_{22}} \end{bmatrix}.$$

Furthermore, $x^\top Ay = \langle A, xy^\top \rangle$ by Q3.3(a). Therefore, its derivative with respect to A is $xy^\top$. Thus, combined together, the final derivative is

$$\begin{bmatrix} \cos(A_{11}^2 + e^{A_{11}+A_{22}})[(2A_{11}) + e^{A_{11}+A_{22}})] + x_1y_1 & x_1y_2 \\ x_2y_1 & \cos(A_{11}^2 + e^{A_{11}+A_{22}}) \cdot e^{A_{11}+A_{22}} + x_2y_2 \end{bmatrix}.$$

[RUBRIC: Taking derivative of $x^\top Ay$ gets **(+1 point)**.]

[RUBRIC: Realizing that $\frac{\partial}{\partial x} f(u(x), v(x)) = \frac{\partial f}{\partial u} \cdot \frac{\partial u}{\partial x} + \frac{\partial f}{\partial v} \cdot \frac{\partial v}{\partial x}$ gets **(+1 point)**.]

[RUBRIC: Correct computation gets **(+1 point)**.]

[RUBRIC: **Total (+3 points)**.]

2. Aside from norms on vectors, we can also impose norms on matrices. Besides the Frobenius norm, the most common kind of norm on matrices is called the induced norm. Induced norms are defined as

$$\|A\|_p = \sup_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p}$$

where the notation $\|\cdot\|_p$ on the right-hand side denotes the vector ℓ_p-norm. Please give the closed-form (or the most simple) expressions for the following induced norms of $A \in \mathbb{R}^{m \times n}$.

(a) $\|A\|_2$. (Hint: Similar to Question 3.4.)

(b) $\|A\|_\infty$.

Solution:

(a) Let $A = U\Sigma V^\top$ be an SVD of A. Then

$$\begin{aligned}
\sup_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} &= \sup_{x\neq 0} \frac{\|U\Sigma V^\top x\|_2}{\|x\|_2} \\
&= \sup_{x\neq 0} \frac{\|\Sigma V^\top x\|_2}{\|x\|_2} \\
&= \sup_{y\neq 0} \frac{\|\Sigma y\|_2}{\|Vy\|_2} (\text{suppose } y = V^\top x) \\
&= \sup_{y\neq 0} \frac{\|\Sigma y\|_2}{\|y\|_2} \\
&= \sigma_1. \text{ (The largest singular value of } A \text{ in absolute value.)}
\end{aligned}$$

(b) We have

$$\begin{aligned}
\sup_{x\neq 0} \frac{\|Ax\|_\infty}{\|x\|_\infty} &= \sup_{x\neq 0} \frac{\max_i(|a_{i1}x_1 + \cdots + a_{in}x_n|)}{\max_i |x_i|} \\
&= \sup_{x\neq 0} \frac{\max_i(|a_{i1}x_1| + \cdots + |a_{in}x_n|)}{\max_i |x_i|} \\
&= \sup_{x\neq 0} \frac{\max_i(|a_{i1}|\,|x_1| + \cdots + |a_{in}|\,|x_n|)}{\max_i |x_i|} \\
&= \max_i \sum_{j=1}^{n} |a_{ij}| \text{ (the largest row sum)}
\end{aligned}$$

[RUBRIC: A complete derivation and correct answer for any of the two sub-part gets **(+1 point)**. **Total (+2 points)**.]

3. (a) Let $\alpha = \sum_{i=1}^{n} y_i \ln \beta_i$ for $y, \beta \in \mathbb{R}^n$. What are the partial derivatives $\frac{\partial \alpha}{\partial \beta_i}$?

(b) Let $\gamma = A\rho + b$ for $b \in \mathbb{R}^n$, $\rho \in \mathbb{R}^m$, $A \in \mathbb{R}^{n\times m}$. What are the the partial derivatives $\frac{\partial \gamma_i}{\partial \rho_j}$?

(c) Given $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$, $z \in \mathbb{R}^k$ and $y = f(x)$, $f : \mathbb{R}^n \mapsto \mathbb{R}^m$, $z = g(y)$, $g : \mathbb{R}^m \mapsto \mathbb{R}^k$. Please write the Jacobian $\frac{dz}{dx}$ as the product of two other matrices. What are these matrices?

(d) Given $x \in \mathbb{R}^n$, $y, z \in \mathbb{R}^m$, and $y = f(x)$, $z = g(x)$. Write the gradient $\nabla_x y^\top z$ in terms of y and z and some other terms.

Solution:

(a) $\frac{\partial \alpha}{\partial \beta_i} = \sum_{j=1}^{n} \frac{\partial (y_j \ln \beta_j)}{\partial \beta_i} = \frac{y_i}{\beta_i}$.

(b) $\frac{\partial \gamma_i}{\partial \rho_j} = A_{ij}$.

(c) Element wise, $\left(\frac{dz}{dx}\right)_{ij} = \frac{dz_i}{dx_j} = \sum_{l=1}^{m} \frac{dz_i}{dy_l}\frac{dy_l}{dx_j}$. Writing this in matrix form gives:

$$\frac{dz}{dx} = \frac{dz}{dy} \cdot \frac{dy}{dx},$$

so $\frac{dz}{dx}$ is the matrix product of other 2 Jacobians.

(d) Again breaking down into element-wise, we have that:

$$\frac{dy^\top z}{dx_i} = \sum_{j=1}^{m} \frac{dy_j}{dx_i} z_j + \frac{dz_j}{dx_i} y_j.$$

Summarizing the above equation into matrix/vector form gives:

$$\nabla_x y^\top z = \left(\frac{dy}{dx}\right)^\top z + \left(\frac{dz}{dx}\right)^\top y.$$

[RUBRIC: A complete derivation and correct answer for any of the 4 sub-part gets **(+1 point)**. **Total (+4 points)**.]

4. Let's apply the multivariate chain rule to a "simple" type of neural network called a *linear neural network*. They're not very powerful, as they can learn only linear regression functions or decision functions, but they're a good stepping stone for understanding more complicated neural networks. We are given an $n \times d$ *design matrix* X. Each row of X is a training point, so X represents n training points with d features each. We are also given an $n \times k$ matrix Y. Each row of Y is a set of k labels for the corresponding training point in X. Our goal is to learn a $k \times d$ matrix W of weights[2] such that

$$Y \approx XW^\top.$$

If n is larger than d, typically there is no W that achieves equality, so we seek an approximate answer. We do that by finding the matrix W that minimizes the *cost function*

$$\mathrm{RSS}(W) = \|XW^\top - Y\|_F^2. \tag{1}$$

This is a classic *least-squares linear regression* problem; most of you have seen those before. But we are solving k linear regression problems simultaneously, which is why Y and W are matrices instead of vectors.

Linear neural networks. Instead of optimizing W over the space of $k \times d$ matrices directly, we write the W we seek as a product of multiple matrices. This parameterization is called a *linear neural network*.

$$W = \mu(W_L, W_{L-1}, \ldots, W_2, W_1) = W_L W_{L-1} \cdots W_2 W_1.$$

Here, μ is called the *matrix multiplication map* (hence the Greek letter mu) and each W_j is a real-valued $d_j \times d_{j-1}$ matrix. Recall that W is a $k \times d$ matrix, so $d_L = k$ and $d_0 = d$. L is the number of *layers* of "connections" in the neural network. You can also think of the network as having $L + 1$ layers of units: $d_0 = d$ units in the *input layer*, d_1 units in the first *hidden layer*, d_{L-1} units in the last hidden layer, and $d_L = k$ units in the *output layer*.

[2] The reason for the transpose on $W^\top$ is because we think in terms of applying W to an individual training point. Indeed, if $X_i \in \mathbb{R}^d$ and $Y_i \in \mathbb{R}^k$ respectively denote the i-th rows of X and Y transposed to be column vectors, then we can write $Y_i \approx WX_i$. For historical reasons, most papers in the literature use design matrices whose rows are sample points, rather than columns.

We collect all the neural network's weights in a *weight vector* $\theta = (W_L, W_{L-1}, \ldots, W_1) \in \mathbb{R}^{d_\theta}$, where $d_\theta = d_L d_{L-1} + d_{L-1} d_{L-2} + \ldots + d_1 d_0$ is the total number of real-valued weights in the network. Thus we can write $\mu(\theta)$ to mean $\mu(W_L, W_{L-1}, \ldots, W_1)$. But you should imagine θ as a column vector: we take all the components of all the matrices $W_L, W_{L-1}, \ldots, W_1$ and just write them all in one very long column vector. Given a fixed weight vector θ, the linear neural network takes an *input vector* $x \in \mathbb{R}^{d_0}$ and returns an *output vector* $y = W_L W_{L-1} \cdots W_2 W_1 x = \mu(\theta) x \in \mathbb{R}^{d_L}$.

Now our goal is to find a weight vector θ that minimizes the composition $\mathrm{RSS} \circ \mu$—that is, it minimizes the cost function

$$J(\theta) = \mathrm{RSS}(\mu(\theta)).$$

We are substituting a linear neural network for W and optimizing the weights in θ instead of directly optimizing the components of W. This makes the optimization problem harder to solve, and you would never solve least-squares linear regression problems this way in practice; but again, it is a good exercise to work toward understanding the behavior of "real" neural networks in which μ is *not* a linear function.

We would like to use a gradient descent algorithm to find θ, so we will derive $\nabla_\theta J$ as follows.

(a) The gradient $G = \nabla_W \mathrm{RSS}(W)$ is a $k \times d$ matrix whose entries are $G_{ij} = \partial \mathrm{RSS}(W)/\partial W_{ij}$, where $\mathrm{RSS}(W)$ is defined by Equation (1). Write two explicit formulas for $\nabla_W \mathrm{RSS}(W)$. First, derive a formula for each G_{ij} using summations, simplified as much as possible. Use that result to find a simple formula for $\nabla_W \mathrm{RSS}(W)$ in matrix notation with no summations.

Solution: Observe that

$$\begin{aligned}
\mathrm{RSS}(W) &= \mathrm{trace}\left(WX^\top XW^\top - WX^\top Y - Y^\top XW^\top + Y^\top Y\right) \\
&= \sum_{p=1}^{k} (WX^\top XW^\top)_{pp} - \sum_{p=1}^{k} (WX^\top Y)_{pp} - \sum_{p=1}^{k} (Y^\top XW^\top)_{pp} + \mathrm{trace}(Y^\top Y) \\
&= \sum_{p=1}^{k} (WX^\top XW^\top)_{pp} - 2\sum_{p=1}^{k} (Y^\top XW^\top)_{pp} + \mathrm{trace}\,(Y^\top Y) \\
&= \sum_{p=1}^{k}\sum_{q=1}^{d}\sum_{r=1}^{n}\sum_{s=1}^{d} W_{pq} X_{rq} X_{rs} W_{ps} - 2\sum_{p=1}^{k}\sum_{q=1}^{n}\sum_{r=1}^{d} Y_{qp} X_{qr} W_{pr} + \mathrm{trace}\,(Y^\top Y).
\end{aligned}$$

Therefore,

$$\begin{aligned}
G_{ij} = \frac{\partial}{\partial W_{ij}} \mathrm{RSS}(W) &= \sum_{q=1}^{d}\sum_{r=1}^{n} W_{iq} X_{rq} X_{rj} + \sum_{r=1}^{n}\sum_{s=1}^{d} X_{rj} X_{rs} W_{is} - 2\sum_{q=1}^{n} Y_{qi} X_{qj} \\
&= 2\left(\sum_{q=1}^{d}\sum_{r=1}^{n} W_{iq} X_{rq} X_{rj} - \sum_{q=1}^{n} Y_{qi} X_{qj}\right)
\end{aligned}$$

This answers the first part of the question. For the second part, note that we can write $G_{ij} = 2(WX^\top X - Y^\top X)_{ij}$, so

$$\nabla_W \text{RSS}(W) = 2(WX^\top - Y^\top)X.$$

[RUBRIC: A complete derivation and correct answer gets **(+1 point)**. Partial answer gets **(+0.5 points)**.]

(b) Directional derivatives are closely related to gradients. The notation $\text{RSS}'_{\Delta W}(W)$ denotes the directional derivative of $\text{RSS}(W)$ in the direction ΔW, and the notation $\mu'_{\Delta\theta}(\theta)$ denotes the directional derivative of $\mu(\theta)$ in the direction $\Delta\theta$.[3] Informally speaking, the directional derivative $\text{RSS}'_{\Delta W}(W)$ tells us how much $\text{RSS}(W)$ changes if we increase W by an infinitesimal displacement $\Delta W \in \mathbb{R}^{k\times d}$. (However, any ΔW we can actually specify is not actually infinitesimal; $\text{RSS}'_{\Delta W}(W)$ is a local linearization of the relationship between W and $\text{RSS}(W)$ at W. To a physicist, $\text{RSS}'_{\Delta W}(W)$ tells us the initial velocity of change of $\text{RSS}(W)$ if we start changing W with velocity ΔW.)

Show how to write $\text{RSS}'_{\Delta W}(W)$ as a Frobenius inner product of two matrices, one related to part (a).

Solution:

$$\text{RSS}'_{\Delta W}(W) = \langle G, \Delta W\rangle_F = \langle 2(WX^\top - Y^\top)X, \Delta W\rangle_F.$$

[RUBRIC: A complete derivation and correct answer gets **(+1 point)**. Partial answer gets **(+0.5 points)**.]

(c) In principle, we could take the gradient $\nabla_\theta\,\mu(\theta)$, but we would need a 3D array to express it! As I don't know a nice way to write it, we'll jump directly to writing the directional derivative $\mu'_{\Delta\theta}(\theta)$. Here, $\Delta\theta \in \mathbb{R}^{d_\theta}$ is a weight vector whose matrices we will write $\Delta\theta = (\Delta W_L, \Delta W_{L-1}, \dots, \Delta W_1)$. Show that

$$\mu'_{\Delta\theta}(\theta) = \sum_{j=1}^{L} W_{>j}\Delta W_j W_{<j}$$

where $W_{>j} = W_L W_{L-1}\cdots W_{j+1}$, $W_{<j} = W_{j-1}W_{j-2}\cdots W_1$, and we use the convention that $W_{>L}$ is the $d_L \times d_L$ identity matrix and $W_{<1}$ is the $d_0 \times d_0$ identity matrix.

Hint: although μ is not a linear function of θ, μ is linear in any *single* W_j; and any directional derivative of the form $\mu'_{\Delta\theta}(\theta)$ is linear in $\Delta\theta$ (for a fixed θ).

Solution: Observe that μ is linear in any single W_j, so

$$\begin{aligned}&\mu(W_L, \dots, W_{j+1}, W_j + \Delta W_j, W_{j-1}, \dots, W_1)\\ &\qquad = \mu(W_L, \dots, W_{j+1}, W_j, W_{j-1}, \dots, W_1) + W_{>j}\Delta W_j W_{<j}.\end{aligned}$$

This linearity implies that

$$\mu'_{\Delta W_j}(\theta) = W_{>j}\Delta W_j W_{<j}.$$

[3] "ΔW" and "$\Delta\theta$" are just variable names that remind us to think of these as small displacements of W or θ; the Greek letter delta is not an operator nor a separate variable.

[RUBRIC: Correct derivation for $\mu'_{\Delta W_j}(\theta)$ gets **(+1 point)**.]

For the specific weight vector $\Delta\theta = (0, \ldots, 0, \Delta W_j, 0, \ldots, 0)$, we have $\mu'_{\Delta\theta}(\theta) = \mu'_{\Delta W_j}(\theta) = W_{>j}\Delta W_j W_{<j}$. But what about for a general weight vector? The directional derivative $\mu'_{\Delta\theta}(\theta)$ is linear in $\Delta\theta$ (for a fixed θ). By linearity, we can just add together the directional derivatives for each ΔW_j, yielding

$$\mu'_{\Delta\theta}(\theta) = \sum_{j=1}^{L} \mu'_{\Delta W_j}(\theta) = \sum_{j=1}^{L} W_{>j}\Delta W_j W_{<j}.$$

[RUBRIC: Correct derivation for final result with the right sum gets **(+1 point)**.]

(d) Recall the chain rule for scalar functions, $\frac{\mathrm{d}}{\mathrm{d}x} f(g(x))|_{x=x_0} = \frac{\mathrm{d}}{\mathrm{d}y} f(y)|_{y=g(x_0)} \cdot \frac{\mathrm{d}}{\mathrm{d}x} g(x)|_{x=x_0}$. There is a multivariate version of the chain rule, which we hope you remember from some class you've taken, and the multivariate chain rule can be used to chain directional derivatives. Write out the chain rule that expresses the directional derivative $J'_{\Delta\theta}(\theta)|_{\theta=\theta_0}$ by composing your directional derivatives for RSS and μ, evaluated at a weight vector θ_0. (Just write the pure form of the chain rule without substituting the values of those directional derivatives; we'll substitute the values in the next part.)

Solution:

$$J'_{\Delta\theta}(\theta)|_{\theta=\theta_0} = \langle \nabla_W \mathrm{RSS}(\mu(\theta)), \mu'_{\Delta\theta}(\theta)\rangle_F|_{\theta=\theta_0}.$$

There are many acceptable variations; e.g.,

$$\begin{aligned}
J'_{\Delta\theta}(\theta)|_{\theta=\theta_0} &= \langle \nabla_W \mathrm{RSS}(\mu(\theta_0)), \mu'_{\Delta\theta}(\theta_0)\rangle_F. \\
J'_{\Delta\theta}(\theta)|_{\theta=\theta_0} &= \langle \nabla_W \mathrm{RSS}(W)|_{W=\mu(\theta)}, \mu'_{\Delta\theta}(\theta)|_{\theta=\theta_0}\rangle_F. \\
J'_{\Delta\theta}(\theta)|_{\theta=\theta_0} &= \langle 2\,(\mu(\theta)\,X^\top - Y^\top)X, \mu'_{\Delta\theta}(\theta)\rangle_F|_{\theta=\theta_0}.
\end{aligned}$$

Note: we will give part marks for $\langle G, \mu'_{\Delta\theta}(\theta)\rangle_F|_{\theta=\theta_0}$, but not full marks unless there is an indication of awareness that G depends on θ_0.

[RUBRIC: Correct derivation gets **(+1 point)**. Any variant of $\langle G, \mu'_{\Delta\theta}(\theta)\rangle_F|_{\theta=\theta_0}$ gets **(+0.5 point)**.]

(e) Now substitute the values you derived in parts (b) and (c) into your expression for $J'_{\Delta\theta}(\theta)$ and use it to show that

$$\begin{aligned}
\nabla_\theta J(\theta) = \; & (2\,(\mu(\theta)\,X^\top - Y^\top)XW_{<L}^\top, \ldots, \\
& 2W_{>j}^\top\,(\mu(\theta)\,X^\top - Y^\top)XW_{<j}^\top, \ldots, \\
& 2W_{>1}^\top\,(\mu(\theta)\,X^\top - Y^\top)X).
\end{aligned}$$

This gradient is a vector in $\mathbb{R}^{d_\theta}$ written in the same format as $(W_L, \ldots, W_j, \ldots, W_1)$. Note that the values $W_{>j}$ and $W_{<j}$ here depend on θ.

Hint: you might find the cyclic property of the trace handy.

Solution:

$$J'_{\Delta\theta}(\theta) = \left\langle 2\,(\mu(\theta)\,X^\top - Y^\top)X, \sum_{j=1}^{L} W_{>j}\Delta W_j W_{<j}\right\rangle_F$$

$$= \sum_{j=1}^{L} \left\langle 2\left(\mu(\theta)\, X^\top - Y^\top\right)X, W_{>j}\Delta W_j W_{<j} \right\rangle_F$$

$$= \sum_{j=1}^{L} \operatorname{trace}\left(2\left(\mu(\theta)\, X^\top - Y^\top\right)X W_{<j}^\top \Delta W_j^\top W_{>j}^\top\right)$$

$$= \sum_{j=1}^{L} \operatorname{trace}\left(2W_{>j}^\top \left(\mu(\theta)\, X^\top - Y^\top\right)X W_{<j}^\top \Delta W_j^\top\right)$$

$$= \sum_{j=1}^{L} \left\langle 2W_{>j}^\top \left(\mu(\theta)\, X^\top - Y^\top\right)X W_{<j}^\top, \Delta W_j \right\rangle_F .$$

As $J'_{\Delta\theta}(\theta) = \nabla_\theta J(\theta) \cdot \Delta\theta$ for all $\Delta\theta \in \mathbb{R}^{d_\theta}$, the gradient $\nabla_\theta J$ must have the value we claim. [RUBRIC: Converting matrix inner product to trace gets **(+0.5 point)**. Application of cyclic property gets **(+0.5 point)**. Rearrangement for final form gets **(+1 point)**.]

5. (**Optional bonus question** worth 1 point. This question contains knowledge that goes beyond the scope of this course, and is intended as an exercise to really make you comfortable with matrix calculus). Consider a differentiable function $f : \mathbb{R}^n \mapsto \mathbb{R}$. Suppose this function admits a unique global optimum $x^* \in \mathbb{R}^n$. Suppose that for some spherical region $\mathcal{X} = \{x \mid \|x - x^*\|^2 \le D\}$ around x^* for some constant D, the Hessian matrix H of the function $f(x)$ is PSD and its maximum eigenvalue is 1. Prove that

$$f(x) - f(x^*) \le \frac{D}{2}$$

for every $x \in \mathcal{X}$. *Hint*: Look up Taylor's Theorem with Remainder. Use Mean Value Theorem on the second order term instead of the first order term, which is what is usually done.

Solution: Start by doing a Taylor expansion at x^* and use mean value theorem:

$$f(x) = f(x^*) + \nabla f(x^*)^\top (x - x^*) + \frac{1}{2}(x - x^*)^\top \nabla^2 f(tx + (1-t)x^*)(x - x^*),$$

where $t \in [0,1]$ is a constant by mean value theorem. We denote $\hat{x} \triangleq tx + (1-t)x^*$, and we know that $\hat{x} \in \mathcal{X}$ since $\hat{x}$ is on the line segment connecting x and x^*, making it also in $\mathcal{X}$ since the region is spherical (In more mathematical language, we say that any convex combination between points in a convex set must lie within that convex set). Therefore we know that $\lambda_{\max}(\nabla^2 f(tx + (1-t)x^*)) = 1$. Thus,

$$(x - x^*)^\top \nabla^2 f(tx + (1-t)x^*)(x - x^*) \le \|x - x^*\|^2.$$

Combining with the fact that $\nabla f(x^*) = 0$ gets:

$$f(x) - f(x^*) \le \frac{1}{2}\|x - x^*\|^2 \le \frac{D}{2}.$$

[RUBRIC: Fully correct gets **(+1 point)**.]

5 Properties of the Normal Distribution (Gaussians)

1. Prove that $\mathbb{E}[e^{\lambda X}] = e^{\sigma^2\lambda^2/2}$, where $\lambda \in \mathbb{R}$ is a constant, and $X \sim \mathcal{N}(0, \sigma^2)$. As a function of λ, $M_X(\lambda) = \mathbb{E}[e^{\lambda X}]$ is also known as the *moment-generating function*.

   Solution:
$$\mathbb{E}[e^{\lambda X}] = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{\lambda x} e^{-x^2/2\sigma^2}dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{\lambda\sigma z}e^{-z^2/2}dz$$
$$= e^{\sigma^2\lambda^2/2}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-(z-\lambda\sigma)^2/2}dz = e^{\sigma^2\lambda^2/2}.$$

   [RUBRIC: A completely correct solution gets **(+1 point)**.]

2. *Concentration inequalities* are inequalities that place upper bounds on the likelihood that a random variable X is far away from its mean μ, written $\mathbb{P}(|X - \mu| \geq t)$, with a falling exponential function ae^{-bt^2} having constants $a, b > 0$. Such inequalities imply that X is very likely to be close to its mean μ. To make a tight bound, we want a to be as small and b to be as large as possible.

   For $t > 0$ and $X \sim \mathcal{N}(0, \sigma^2)$, prove that $\mathbb{P}(X \geq t) \leq \exp(-t^2/2\sigma^2)$, then show that $\mathbb{P}(|X| \geq t) \leq 2\exp(-t^2/2\sigma^2)$.
   Hint: Consider using Markov's inequality and the result from Question 5.1.

   Solution: For any $\lambda > 0$,
$$\mathbb{P}(X \geq t) = \mathbb{P}(e^{\lambda X} \geq e^{\lambda t}) \leq e^{-\lambda t}\mathbb{E}[e^{\lambda X}] = e^{-\lambda t}e^{\sigma^2\lambda^2/2},$$
   where the inequality applies Markov's inequality. Setting $\lambda = t/\sigma^2$ gives the claim
$$\mathbb{P}(X \geq t) \leq \exp(-t^2/2\sigma^2).$$
   Due to the symmetry of X we can then conclude that
$$\mathbb{P}(|X| \geq t) \leq 2\exp(-t^2/2\sigma^2).$$

   [RUBRIC: A completely correct solution gets **(+1 point)**.]

3. Let $X_1, \dots, X_n \sim \mathcal{N}(0, \sigma^2)$ be i.i.d. (independent and identically distributed). Find a concentration inequality, similar to Question 5.2, for the average of n Gaussians: $\mathbb{P}(\frac{1}{n}\sum_{i=1}^n X_i \geq t)$? What happens as $n \to \infty$?
   Hint: Without proof, use the fact that linear combinations of i.i.d. Gaussian-distributed variables are also Gaussian-distributed. Be warned that summing two Gaussian variables does **not** mean that you can sum their probability density functions (no no no!).

   Solution: From the hint we know that $\frac{1}{n}\sum_{i=1}^n X_i$ follows a Gaussian distribution, so we only need to determine its mean and variance. Its mean is clearly 0. Its variance is
$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\, n\mathrm{Var}(X_i) = \frac{\sigma^2}{n},$$

where we use the fact that the variance of a sum of uncorrelated variables separates into a sum of their variances.

[RUBRIC: Correct mean value gets **(+0.5 point)**.]

[RUBRIC: Correct variance value gets **(+0.5 point)**.]

Now we apply the concentration result of the previous part to conclude that

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq t\right) \leq \exp(-nt^2/2\sigma^2).$$

[RUBRIC: A correct concentration result gets **(+1 point)**.]

As $n \to \infty$, the probability of the empirical mean $\frac{1}{n}\sum_{i=1}^{n} X_i$ not being zero vanishes. This is usually expressed as $\frac{1}{n}\sum_i X_i$ "converges in probability" to the constant 0. The phenomena that the empirical mean of i.i.d. random variables $\frac{1}{n}\sum_i X_i$ (not necessarily Gaussian) converges in probability to its true mean is called the *Weak Law of Large Numbers*.

[RUBRIC: Realizing that probability of empirical mean *being non-zero* converges to zero gets **(+1 point)**. Caution: probability of being a given non-zero value is always zero as it is a continuous variable.]

[RUBRIC: **Total (+3 points)**.]

4. Let vectors $u, v \in \mathbb{R}^n$ be constant (i.e., not random) and orthogonal (i.e., $\langle u, v\rangle = u \cdot v = 0$). Let $X = (X_1, \ldots, X_n)$ be a vector of n i.i.d. standard Gaussians, $X_i \sim \mathcal{N}(0, 1), \forall i \in [n]$. Let $u_x = \langle u, X\rangle$ and $v_x = \langle v, X\rangle$. Are u_x and v_x independent? Explain. If $X_1, \ldots, X_n$ are independently but not identically distributed, say $X_i \sim \mathcal{N}(0, i)$, does the answer change? *Hint*: two jointly normal random variables are independent if and only if they are uncorrelated.

Solution: As u_x and v_x are two different linear transformations of a normally distributed random vector, they are jointly normal random variables. Hence they are independent if and only if they are uncorrelated. The covariance of u_x and v_x is

$$\mathbb{E}[u_x v_x] = \mathbb{E}\left[\left(\sum_{i=1}^{n} u_i X_i\right)\left(\sum_{i=1}^{n} v_i X_i\right)\right] = \sum_{i=1}^{n} u_i v_i \mathbb{E}[X_i^2] = \langle u, v\rangle = 0.$$

[RUBRIC: Correct argument for u_x and v_x being independent for i.i.d. X_i gets **(+1 point)**.]

Therefore, u_x and v_x are independent. However, if $X_1, \ldots, X_n$ are not identically distributed, $\mathbb{E}[u_x v_x] = \sum_{i=1}^{n} u_i v_i \mathbb{E}[X_i^2] = \sum_{i=1}^{n} u_i v_i i$, not necessarily equal to 0. Therefore if the X_i's are not identically distributed, u_x and v_x are not necessarily independent.

[RUBRIC: Correct derivation and value of covariance between u_x and v_x when $X_i \sim \mathcal{N}(0, i)$ gets **(+0,5 point)**.]

[RUBRIC: Arguing that u_x and v_x may not independent for non-iid X_i gets **(+0.5 point)**.]

[RUBRIC: **Total (+2 points)**.]

6 The Multivariate Normal Distribution

The multivariate normal distribution with mean $\mu \in \mathbb{R}^d$ and positive definite covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$, denoted $\mathcal{N}(\mu, \Sigma)$, has the probability density function

$$f(x; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right).$$

Here $|\Sigma|$ denotes the determinant of Σ. You may use the following facts without proof.

- The volume under the normal PDF is 1.
$$\int_{\mathbb{R}^d} f(x)\,\mathrm{d}x = \int_{\mathbb{R}^d} \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp\left\{-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)\right\}\mathrm{d}x = 1.$$
- The change-of-variables formula for integrals: let f be a smooth function from $\mathbb{R}^d \to \mathbb{R}$, let $A \in \mathbb{R}^{d\times d}$ be an invertible matrix, and let $b \in \mathbb{R}^d$ be a vector. Then, performing the change of variables $x \mapsto z = Ax + b$,
$$\int_{\mathbb{R}^d} f(x)\,\mathrm{d}x = \int_{\mathbb{R}^d} f(A^{-1}z - A^{-1}b)\,|A^{-1}|\,\mathrm{d}z.$$

1. Let $X \sim \mathcal{N}(\mu, \Sigma)$. Use a suitable change of variables to show that $\mathbb{E}[X] = \mu$.

   Solution: With the change of variables $x \mapsto z = x - \mu$,
$$\begin{aligned}\mathbb{E}[X] &= \int_{\mathbb{R}^d} \frac{x}{\sqrt{(2\pi)^d|\Sigma|}} \exp\left\{-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)\right\}\mathrm{d}x\\ &= \int_{\mathbb{R}^d} \frac{\mu + z}{\sqrt{(2\pi)^d|\Sigma|}} \exp\left\{-\frac{1}{2}z^\top\Sigma^{-1}z\right\}\mathrm{d}z\\ &= \mu\int_{\mathbb{R}^d} f(z; 0, \Sigma)\,\mathrm{d}z + \int_{\mathbb{R}^d} z f(z; 0, \Sigma)\,\mathrm{d}z = \mu + 0 = \mu.\end{aligned}$$

   The left integral contains the normal PDF, whose integral is 1. The right integral integrates to 0 because $zf(z; 0, \Sigma)$ is an odd function: the contributions of z and $-z$ cancel each other out.

   [RUBRIC: Total **(+2 points)**]

2. Use a suitable change of variables to show that $\mathrm{Var}(X) = \Sigma$, where the variance of a vector-valued random variable X is
$$\mathrm{Var}(X) = \mathrm{Cov}(X, X) = \mathbb{E}[(X-\mu)(X-\mu)^\top] = \mathbb{E}[XX^\top] - \mu\mu^\top.$$

   Hints: Every symmetric, positive semidefinite matrix Σ has a symmetric, positive definite square root $\Sigma^{1/2}$ such that $\Sigma = \Sigma^{1/2}\Sigma^{1/2}$. Note that Σ and $\Sigma^{1/2}$ are invertible. After the change of variables, you will have to find another variance $\mathrm{Var}(Z)$; if you've chosen the right change of variables, you can solve that by solving the integral for each diagonal component of $\mathrm{Var}(Z)$

and a second integral for each off-diagonal component. The diagonal components will require integration by parts.

Solution: With the change of variables $x \mapsto z = \Sigma^{-1/2}(x - \mu)$,

$$\begin{aligned}
\mathrm{Var}(X) &= \int_{\mathbb{R}^d} (x-\mu)(x-\mu)^\top \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left\{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right\} \mathrm{d}x \\
&= \int_{\mathbb{R}^d} \Sigma^{1/2} z z^\top \Sigma^{1/2} \frac{1}{\sqrt{(2\pi)^d}} \exp\left\{\frac{1}{2} z^\top z\right\} \mathrm{d}z \\
&= \mathbb{E}_{Z\sim\mathcal{N}(0,I)}[\Sigma^{1/2} Z Z^\top \Sigma^{1/2}] = \Sigma^{1/2} \mathbb{E}_{Z\sim\mathcal{N}(0,I)}[Z Z^\top] \Sigma^{1/2} \\
&= \Sigma^{1/2} \mathrm{Var}(Z) \Sigma^{1/2}.
\end{aligned}$$

[RUBRIC: Correct derivation of the decomposition of Var(X) **(+1 point)**]

Thus, if we can show that $\mathrm{Var}(Z) = I$ where $Z \sim \mathcal{N}(0, I)$, we are done. To that end, we can compute the diagonal and off-diagonal components separately:

- For $i \neq j$, we have

$$\int_{\mathbb{R}^d} z_i z_j \frac{1}{\sqrt{(2\pi)^d}} e^{-\frac{1}{2} z^\top z} \mathrm{d}z = \frac{1}{2\pi} \int_{-\infty}^{\infty} z_i e^{-\frac{1}{2} z_i^2} \mathrm{d}z_i \int_{-\infty}^{\infty} z_j e^{-\frac{1}{2} z_j^2} \mathrm{d}z_j = 0.$$

- For $i = j$, we have

$$\int_{\mathbb{R}^d} z_i^2 \frac{1}{\sqrt{(2\pi)^d}} e^{-\frac{1}{2} z^\top z} \mathrm{d}z = \int_{-\infty}^{\infty} z_i^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} z_i^2} \mathrm{d}z_i = \left[-z_i \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} z_i^2}\right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} z_i^2} \mathrm{d}z_i = 1$$

  where we used integration by parts in the second equality.

This suffices to prove that $\mathrm{Var}(Z) = I$. Therefore, $\mathrm{Var}(X) = \Sigma$.

[RUBRIC: Correct derivation of Var(Z) **(+1 point)**]

[RUBRIC: Total **(+2 point)**]

7 Gradient Descent

Consider the optimization problem $\min_{x \in \mathbb{R}^n} \frac{1}{2} x^\top A x - b^\top x$, where $A \in \mathbb{R}^{n \times n}$ is a PSD matrix with $0 < \lambda_{\min}(A) \le \lambda_{\max}(A) < 1$.

1. Find the optimizer x^* (in closed form).

   Solution: Since the objective is convex, the optimizer is a stationary point of the objective, i.e., it satisfies
   $$Ax - b = 0,$$
   and since A is invertible, the optimizer is $x^* = A^{-1} b$.

   [RUBRIC: A correct derivation and answer gets **(+1 point)**.]

2. Solving a linear system directly using Gaussian elimination takes $O(n^3)$ time, which may be wasteful if the matrix A is sparse. For this reason, we will use gradient descent to compute an approximation to the optimal point x^*. Write down the update rule for gradient descent with a step size of 1 (i.e., taking a step whose length is the length of the gradient).

   Solution: $x^{(k+1)} = x^{(k)} - (Ax^{(k)} - b)$.

   [RUBRIC: Correct update law gets **(+1 point)**.]

3. Show that the iterates $x^{(k)}$ satisfy the recursion $x^{(k)} - x^* = (I - A)(x^{(k-1)} - x^*)$.

   Solution: We expand the gradient descent update to obtain
   $$\begin{aligned} x^{(k)} - x^* &= x^{(k-1)} - (Ax^{(k-1)} - b) - x^* = (I - A)x^{(k-1)} + b - x^* \\ &= (I - A)x^{(k-1)} - (I - A)x^* = (I - A)(x^{(k-1)} - x^*). \end{aligned}$$
   In the third equality we used the stationary condition $Ax^* = b$.

   [RUBRIC: Correct argument gets **(+1 point)**.]

4. Using Question 3.4, prove $\|Ax\|_2 \le \lambda_{\max(A)} \|x\|_2$.
   Hint: Use the fact that, if λ is an eigenvalue of A, then λ^2 is an eigenvalue of A^2.

   Solution: We can write $\|Ax\|_2^2 = x^\top A^2 x$. First assume x has unit length. Then,
   $$\|Ax\|_2^2 = x^\top A^2 x \le (\lambda_{\max}(A))^2.$$
   Now take any $x \neq 0$, not necessarily of unit length ($x = 0$ trivially satisfies the inequality). Then, we have proved that
   $$\|A(x/\|x\|_2)\|_2^2 \le (\lambda_{\max}(A))^2.$$
   Multiplying both sides by $\|x\|_2^2$ and taking the square root completes the proof of the identity.

   [RUBRIC: Correct argument gets **(+1 point)**.]

5. Using the previous two parts, show that for some $0 < \rho < 1$,

$$\|x^{(k)} - x^*\|_2 \leq \rho \|x^{(k-1)} - x^*\|_2.$$

Solution: Note that $I - A \succ 0$, because $\lambda_{\max}(A) < 1$. Therefore

$$\|(I - A)(x^{(k-1)} - x^*)\|_2 \leq \lambda_{\max}(I - A)\|x^{(k-1)} - x^*\|_2.$$

Let $\rho = \lambda_{\max}(I - A) = 1 - \lambda_{\min}(A)$, which is in $(0, 1)$ because $\lambda_{\min}(A) \leq \lambda_{\max}(A) < 1$ and $\lambda_{\min}(A) > 0$. Then

$$\|x^{(k)} - x^*\|_2 = \|(I - A)(x^{(k-1)} - x^*)\|_2 \leq \rho \|x^{(k-1)} - x^*\|_2.$$

[RUBRIC: Correct argument gets **(+1 point)**.]

6. Let $x^{(0)} \in \mathbb{R}^n$ be the starting value for our gradient descent iterations. If we want a solution $x^{(k)}$ that is $\epsilon > 0$ close to x^*, i.e. $\|x^{(k)} - x^*\|_2 \leq \epsilon$, then how many iterations of gradient descent should we perform? In other words, how large should k be? Give your answer in terms of $\rho, \|x^{(0)} - x^*\|_2$, and ϵ.

Solution: Unrolling the recursion of part (d) gives us

$$\|x^{(k)} - x^*\|_2 \leq \rho^k \|x^0 - x^*\|_2.$$

[RUBRIC: Correct application of recursion for $\|x^{(k)} - x^*\|_2$ gets **(+1 point)**.]

Therefore a sufficient condition for $\|x^{(k)} - x^*\|_2 \leq \epsilon$ to hold true is

$$\rho^k \|x^0 - x^*\|_2 \leq \epsilon.$$

Taking logarithms and rearranging, this yields

$$k \geq \frac{1}{\log \frac{1}{\rho}} \log\left(\frac{\|x^0 - x^*\|_2}{\epsilon}\right).$$

[RUBRIC: Correct inequality for k gets **(+1 point)**.]

[RUBRIC: **Total (+2 points)**.]